# Robust calibrations on reduced sample sets for API content prediction in tablets: Definition of a cost-effective NIR model development strategy

Sigrid Pieters[1], Wouter Saeys[2], Tom Van den Kerkhof[3], Mohammad Goodarzi[2], Mario Hellings[3], Thomas De Beer[4], Yvan Vander Heyden[1]

[1] Department of Analytical Chemistry and Pharmaceutical Technology, Center for Pharmaceutical Research, Vrije Universiteit Brussel - VUB, Laarbeeklaan 103, B-1090 Brussels, Belgium
[2] Department of Biosystems, Faculty of Bioscience Engineering, Katholieke Universiteit Leuven – KULeuven, Kasteelpark Arenberg 30, B-3001 Heverlee, Belgium
[3] ChemPhar development Johnson & Johnson, Turnhoutseweg 30, B-2340 Beerse, Belgium
[4] Laboratory of Pharmaceutical Process Analytical Technology, Universiteit Gent, Harelbekestraat 72, B-9000 Ghent, Belgium

* Corresponding author: Tel.: +32 2 477 47 34
Fax: +32 2 477 47 35
Email: yvanvdh@vub.ac.be (Y. Vander Heyden)

Keywords: near infrared spectroscopy, prior spectral information, tablets, orthogonal projection methods, augmentation methods

**Abstract**

Owing to spectral variations from other sources than the component of interest, large investments in the NIR model development may be required to obtain satisfactory and robust prediction performance. To make the NIR model development for routine active pharmaceutical ingredient (API) prediction in tablets more cost-effective, alternative modelling strategies were proposed. They used a massive amount of prior spectral information on intra- and inter-batch variation and the pure component spectra to define a clutter, i.e. the detrimental spectral information. This was subsequently used for artificial data augmentation and/or orthogonal projections. The model performance improved statistically significantly, with a 34-40% reduction in RMSEP while needing fewer model latent variables, by applying the following procedure before PLS regression: *(1)* augmentation of the calibration spectra with the spectral shapes from the clutter, and *(2)* net analyte pre-processing (NAP). The improved prediction performance was not compromised when reducing the variability in the calibration set, making exhaustive calibration unnecessary. Strong water content variations in the tablets caused frequency shifts of the API absorption signals that could not be

included in the clutter. Updating the model for this kind of variation demonstrated that the completeness of the clutter is critical for the performance of these models and that the model will only be more robust for spectral variation that is not co-linear with the one from the property of interest.

## 1. Introduction

Together with the growing popularity of near-infrared (NIR) spectroscopy, the interest to use multivariate prediction models (e.g. PLS) emerged in different industries (food, chemical, pharmaceutical…) [1-3]. Compared to HPLC, NIR spectroscopy provides a non-destructive, extremely fast, and hands-on analysis, which can substantially reduce analysis times and costs for the routine assay of active pharmaceutical ingredient (API) in tablets. Multivariate calibration models (e.g. PLS) are typically built in an empirical way. Although this may offer flexibility, it can also be a drawback making the model development, maintenance and update more complex [4]. Consequently, more investment may be required, which must guarantee that a low prediction error is obtained and maintained for future predictions of novel batches.

There are numerous examples where spectral disturbances from various sources, e.g. variations in a process, temperature, humidity... , upset the model predictions. In this paper, we specifically consider this problem in tablets produced from a routine production line. Besides measurement variability (path length, temperature...), systematic changes in the measured signals in the spectra of tablets may appear, e.g. originating from process and raw material variability (inter- and intra-batch effects). A tableting process involves different subsequent processing steps, such as mixing the ingredients, granulation, compaction and coating. Some tablet production variables, e.g. grain size, compaction pressure, and coating thickness, may introduce spectral variability that is irrelevant for predicting the API content in the tablets [5]. Also, the excipients in the tablet can vary slightly between different batches or samples (excipient homogeneity), while the tablet composition can vary as well when different API target concentration levels are considered in the model. Depending on the relative humidity conditions during their production, packaging, storage and analysis, the water content is another important variable that may upset the NIR model predictions [6].

It is evident that when more relevant variation is covered by the calibration set of the NIR model, lower prediction errors can be obtained on new independent samples [7]. Including all the expected variation in the calibration set is the most intuitive way for doing this, and is still recommended by regulatory authorities [8]. Drawbacks of this approach are that the model complexity may increase rapidly when more 'non-relevant' spectral variability is included in the

exhaustive calibration, needing additional latent variables or introducing non-linearities that may require non-linear calibration techniques. A more important practical problem is that a substantial number of well-chosen samples, describing all the expected variability, is needed to obtain robust model performance. It is a non-trivial task to obtain a good estimate of the expected variation in the samples to be predicted, also because there is a practical limit (particularly in terms of costs) to the number of samples to be analysed with the reference technique. Hence, in most practical cases one selects randomly the calibration samples from 'representative' batches. As a result, existing calibration bases rarely contain all the relevant variability from all the influence factors which can occur in industrial conditions [9], making the models less robust for their long term use and requiring frequently model updates.

It is often overlooked that one of the major advantages of NIR spectroscopy, i.e. its analysis speed and simplicity, provides an excellent opportunity to measure significantly more dosage units than what would be possible with the reference technique, i.e. HPLC. This is especially the case when automated NIR equipment is available. Hence, a massive amount of spectral information can be obtained easily to allow a better understanding of the possible spectral perturbations. The aim of the present study is to investigate whether the use of such spectral information on intra- and inter-batch variation (without corresponding reference analyses) during model development can improve the NIR model performance for predicting the API content in tablets from novel batches. The prediction performance of different approaches using the prior spectral information was compared to that of PLS models. It was also investigated whether these strategies allowed reducing the variability in the calibration set without compromising the prediction performance. This should lead to the definition of a cost-effective strategy for developing robust calibration models for the routine prediction of API in tablets.

## 2. Theory

There are different ways to incorporate prior information into the calibration model. One distinguishes augmentation and orthogonal correction methods (supervised and unsupervised), while combinations of these are possible as well.

### 2.1. Augmentation methods

#### *2.1.1. Noise augmentation in PLS*

The idea of noise augmentation (NA) or ensemble methods [10-13] is to expand the calibration set artificially in order to build PLS models that are more robust to the different kinds of

variability expected in the future sample population. An artificial calibration set can be created by augmenting the original one with a high number of spectral signatures, representing the 'noise'. The original spectra in the resulting calibration set should span the chemical variation in the best possible way, whereas the perturbation spectra should represent all the other possible variations that can be expected. One way to estimate the latter is through the measurement of spectra under varying perturbation conditions. Because the 'noise' added to the calibration spectra should be independent from the component of interest, this operation should not change the corresponding reference values.

*2.1.2. Prediction – Augmented Classical Least Squares (P-ACLS)*

In contrast to the inverse modeling approaches (e.g. PLS), classical least squares (CLS) regression is based on an explicit linear additive model (e.g. Lambert-Beer's law in spectroscopy). Equation (1) depicts the CLS model.

$$\mathbf{J} = \mathbf{PK} + \mathbf{E_A} \quad (1)$$

where **J** is the matrix of the measured intensities, **P** is the matrix of concentration values, **K** is the matrix of the pure component signals at unit concentration, and $\mathbf{E_A}$ the model error.

The major weakness of CLS is that it requires quantitative knowledge of all the spectrally active components in the calibration set to get an estimate of **K**. Augmented (A)-CLS attempts to obtain a better estimation of **K** by augmenting the predicted pure component matrix with empirically derived spectral shapes, e.g. the loading vectors from PCA on the spectral residuals $\mathbf{E_A}$ (SRACLS), known pure component spectra [14], or a priori known other variation not included in the calibration set (PACLS) [15-16]. The addition of the spectral shapes both changes and corrects the concentration estimates for the component of interest.

**2.2. Orthogonal correction methods**

*2.2.1. Orthogonal projections*

There are essentially 2 contributions within the calibration matrix **X**, i.e. one originating from the analyte of interest *k* ($\mathbf{X}_k$), and another from all other sources of variance ($\mathbf{X}_{-k}$).

$$\mathbf{X} = \mathbf{X}_k + \mathbf{X}_{-k} \quad (2)$$

Pretreatments based on orthogonal projections aim at removing those spectral patterns, which are 'interfering' with the desired prediction from the data matrix **X**, before calibration on **y**. It is attempted to return the spectra of **X** as **X***, containing the most condensed spectral information, i.e. the net analyte signal ($\mathbf{X}_k$) [17]. The information to be removed, i.e. the clutter, can be defined based on pure component spectra of known interferences [18], an experimental design with varying perturbation factor(s) [19], or an (augmented) calibration data set [12]. The column vectors of the so obtained matrix **S** form a basis of the detrimental subspace. Then, an orthogonal projector to **S**, i.e. $P_S^{\perp}$, can be calculated to correct the initial matrix **X** as follows.

$$\mathbf{X}^* = \mathbf{X}P_S^{\perp} = \mathbf{X}(\mathbf{I} - \mathbf{S}(\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}) \qquad (3)$$

*2.2.1.1. External Parameter Orthogonalisation (EPO)*

This unsupervised orthogonalisation method uses pure spectra (without reference value) to define a basis of the space spanned by the 'interfering' factors, this way estimating the parasitic subspace $\mathbf{X}_k$ [19]. PCA is applied to **D**, containing spectra collected while the perturbation factor is varying. Retaining only the first $g_a$ PC's, the column vectors of the matrix of eigenvectors **G** will represent an orthonormal basis of the subspace to be removed. Subsequently, an orthogonal projection is defined to filter the calibration spectra **X** to obtain the 'corrected' ones (**X***).

$$\mathbf{X}^* = \mathbf{X}(\mathbf{I} - \mathbf{G}\mathbf{G}^{\mathrm{T}}) \qquad (4)$$

where G is a matrix comprising the $g_a$ first eigenvectors of the square matrix $\left[\mathbf{D}^{\mathrm{T}}\mathbf{D}\right]$.

Because the original calibration data base X is adjusted by means of orthogonal projection, the correction is embedded into the model. Hence, there is no need to reapply the correction to new spectra when using the model. Advantages of this method are the high flexibility and the fact that it does not require corresponding y values. Because it is based on external a priori information, the success of the unsupervised orthogonalisation will strongly depend on how one is able to identify the detrimental spaces within the variables space, without interfering with the useful space. It will depend on the comprehensiveness of the a priori spectral information and on the empirical tuning for identifying the information to be removed. In practice, the latter can be supported by assessing the RMSECV as a function of $g_a$ and the number of latent variables, or by the evolution of the Wilks lambda value [19]. Nevertheless, using unsupervised orthogonal projection methods holds a risk of

removing too much information, i.e. when there is detrimental information not independent with the net analyte signal [20].

*2.2.1.2. Orthogonal Signal Correction (OSC)*

OSC intends to subtract those factors from the calibration matrix **X** which capture the variability in **X** orthogonal to **y** (supervised orthogonalisation) [21]. It uses a PLS-NIPALS-like algorithm, where the weighted regression vector **w** is adjusted. Thus, OSC pre-processing involves two major steps, i.e. (1) estimation of the vector pair (**t,w**) for which $\mathbf{tw}^{\mathbf{T}}$ explains maximum variance in **X** orthogonal to **y**, and (2) removal of the contribution of the identified component from **X**.

$$\mathbf{X}^{*} = (\mathbf{X} - \mathbf{tp}^{\mathbf{T}}) \qquad (5)$$

Then, the newly obtained **X*** can be calibrated on **y** by means of PLS. Different approaches have been proposed for estimating the score vector **t**. As the orthogonal projection is not embedded in the calibration model, it has to be applied to new spectra prior to applying the calibration model [17].

*2.2.1.3. Net Analyte Pre-processing (NAP)*

This method was introduced by Goicoechea and Olivieri [22]. To obtain an estimate of the parasitic subspace $X_{-k}$, the following orthogonal projection of X to y is performed.

$$\mathbf{X}_{-k}est = (\mathbf{I} - \mathbf{y}(\mathbf{y}^{\mathbf{T}}\mathbf{y})^{-1}\mathbf{y}^{\mathbf{T}})\mathbf{X} \qquad (6)$$

In a next step, **X** is projected orthogonal to the matrix **U**, consisting of the first $g_a$ PC's of $\mathbf{X}_{-k}est$ to obtain **X*** for calibration on **y** by means of PLS or CLS.

$$\mathbf{X}^{*} = \mathbf{X}(\mathbf{I} - \mathbf{UU}^{\mathbf{T}}) \qquad (7)$$

The same transformation is applied to new spectra *x* prior to applying the calibration model.

$$x^{*} = (\mathbf{I} - \mathbf{UU}^{\mathbf{T}})x \qquad (8)$$

Thus, NAP has in common with OSC that it attempts removing the spectral information in **X** that is orthogonal to **y** (supervised orthogonalisation methods). Yet both methods use different routes to reach this goal. Compared to unsupervised orthogonalisation methods, OSC and NAP allow removing

the detrimental subspace in a less empirical way, and may also work efficiently when detrimental and useful information are not independent [20]. A disadvantage is that all the variation to be removed should be present in the calibration base, thus corresponding y values are necessary. Because of their close relationship to PLS, it has been repeatedly demonstrated that these methods can reduce the complexity of the PLS regression model, but do not enhance its predictive power [22-24].

## 3. Experimental

### 3.1. NIR spectroscopy

Tablets were analyzed using a FT-NIR spectrometer (MPA, Bruker, Ettlingen, Germany). Spectra were collected in the 10000-5700 $cm^{-1}$ region with a resolution of 8 $cm^{-1}$ and averaged over 16 scans. The effective sample size was approximately 11% of the tablet. It was assumed that the API was uniformly distributed in the tablet, which was confirmed by studies during method development.

### 3.2. HPLC

HPLC was used as the reference technique. All HPLC measurements were performed at Janssen Pharmaceutica, Beerse, Belgium, using their developed and validated method (confidential).

### 3.3. Data analysis

PCA, PLS regression and OSC pre-treatment were performed in Matlab 7.1. (The Mathworks, Nattick, MA) using the PLS toolbox 6.2. (Eigenvector Research, Wenatchee, WA). EPO, ACLS and NAP were directly programmed in Matlab. Minitab 16 (Minitab, PA) was used for testing the statistical significance of the predictions by the different models.

#### *3.3.1. Model development and validation*

The calibration set consisted of oblong shaped tablets of 6 target API concentration levels, i.e. 0, 25, 50, 100, 150 and 187.5 mg API with respective 0, 7.28, 14.56, 29.12, 38.83 and 48.54% $ww^{-1}$ per tablet. Tablets at the extreme concentration levels originated from laboratory batches, and were manufactured to extend the concentration range spanning the chemical variability within the specification limits. The tablets at the other concentration levels were manufactured in a production line and represent tablets to be marketed. As the calibration set contained a large concentration range, the tablet composition also varied over the concentration levels.

The NIR spectra were limited to the 9000-7600 $cm^{-1}$ range and were pre-processed with standard normal variates (SNV). Except for the ACLS and NAP-CLS models, all other calibration models were built with PLS. Mean-centering was always performed prior to PLS modelling. The optimal number of latent variables for each model was selected based on the minimal RMSECV from 'leave-one-batch-out' cross validation. Two independent test sets were used for external model validation. Test set 1 consisted of 54 tablets, originating from 4 new production batches (one for each dose to be marketed) and 2 laboratory batches at the extreme concentration levels. In test set 2 (30 tablets from 4 production batches) strong water variations were introduced by storing the tablets at different conditions (see further in 3.3.2.). The root mean square error of prediction (RMSEP) was used as the performance criterion to assess and compare the predictive abilities of the different models. The significance of differences in prediction power was assessed by a two-way ANOVA, performed on the absolute values of the prediction errors [25]. Multiple comparisons were performed by the Dunnett's test, using the PLS model as a control to calculate *p* values.

*3.3.2. Available prior information*

NIR spectra were recorded from the pure tablet components, i.e. from the API (**A**) and the excipients (**B**). Fig.1 shows the obtained spectra in the considered spectral range after SNV pre-processing. 10086 NIR spectra from tablets originating from 27 different production batches (i.e. for doses to be marketed: 3 batches at 7.28 % w/w, 12 batches at 14.56 %w/w, 7 batches at 29.12 % w/w and 5 batches at 38.83 % w/w) were measured. They included tablets produced during pharmaceutical development and clinical trials. They also included characterization batches, produced at the limits of the critical process settings, to increase the spectral variability originating from process variability. Since hundreds of tablets per batch were measured, systematic variations in the spectra due to process variation over time may also be covered. The NIR spectra of the tablets originating from laboratory batches (at the extreme target API concentration levels) were added to this matrix, creating **C** (10213 x 364). 107 spectra from processed tablets with deliberate strong water variations were measured (Fig.2). Either the tablets were stored for 30 min at 50°C (to decrease the water content in the tablets), either for 3h at 75% RH, or for 16h at 75% RH (to increase the water content in the tablets).

**3.4. Strategies for efficient use of the prior information**

*3.4.1. Identifying the information to be removed (clutter)*

The strategies applied in this paper do not require exact knowledge of all perturbation sources, but just need a representative estimate of the spectral variation to be expected in tablets

from future batches. This can be obtained by making use of the extremely fast and simple NIR analysis, where far more tablets could be analyzed than what would be possible with HPLC. Fig.3 shows a PCA score plot (PC1 versus PC2) of samples from randomly selected batches (coloured markers) for calibration of the initial PLS model, and all the prior spectral information **C** (10213 x 364) are in grey dots. The scores for **C** indicate a generally higher spread for the production batches, and this was confirmed for the following investigated PC's as well (not shown). The PCA score plots of the spectra of **C** from tablets at each target concentration level also showed clear systematic effects between and within production batches. Examples are shown in Figs. 4-5. An interesting example of intra-batch variability is shown in Fig.4, where the spectra from the batch marked with blue squares showed two separate clusters. This may be due to a systematic effect that occurred over the manufacturing time of the batch. Hence, selecting the samples for calibration holds a risk of not fully covering the spectral intra- and inter-batch variation that may be present in future samples.

As the variation in **C** is partly due to differences in API content (the parameter to be quantified) in each individual tablet, this spectral information cannot be used 'as such' for defining the clutter for the proposed strategies. Many correction methods define the clutter by the difference spectra at various perturbation levels, which works well when the property to be determined can be held constant [17]. Here, it is impossible to obtain individual tablets, even from the same batch, with exactly the same API concentration and tablet composition. Another possibility is using a y-gradient method to select the spectra for calculating the difference spectra [26], but no y-reference values are available for the massive amount of prior spectral information. For the present case study, we proposed another methodology to obtain a relevant estimate of the clutter. To correct **C** for known chemical variations, the pure spectra from the tablet constituents were used to calculate a basis. Removal of their spectral contributions from **C** was performed by orthogonal projection [18]. After the correction of **C**, the spectra have lost their initial form, leaving only the spectral shapes from other variations (Fig.6).

Fig.7 schematically illustrates the different tested strategies for obtaining a clutter. Matrix **B** (tablet excipients) can be defined as a clutter containing the known chemical noise. Considering both **A** and **B** for the orthogonal correction renders a clutter containing only the physical variations and unknown chemical variation (e.g. water content...) (**D**). Using **A** yields a clutter containing physical variations and chemical variability caused by all 'interfering' excipients (**E**).

*3.4.2. Strategies using EPO*

Depending on how the clutter is defined, there are different possibilities to perform EPO corrections on the calibration spectra **X** (Fig.7). In a similar way as above, $EPO_{chem}$ directly removes the known interfering chemical variations (**B**) from the **X** spectra. In $EPO_{glob}$ the expected chemical

and other variations were captured in **E** for global EPO correction of **X**. Here, optimization of the dimension of the subspace to be withdrawn was performed by analyzing the evolution of the RMSECV as a function of $g_a$ and the number of LV's [19].

*3.4.3. Strategies using data augmentation*

Noise augmentation (NA) of the calibration set was obtained by adding the mean-centered spectra of **D** to *n* repetitions of the original calibration spectra **X**. Similarly, $\mathbf{y}_{aug}$ was developed as *n* repetitions of **y**, because the added spectral variations are not supposed to change the corresponding *y* value of each spectrum (Fig.7). Hence, a calibration set containing 10213 objects was obtained and used for PLS regression.

Because OSC and NAP need a calibration data base to calculate the orthogonal projection, these techniques were applied on the augmented calibration set, i.e., $\mathbf{X}_{aug}$ and $\mathbf{y}_{aug}$ containing the most complete information. The $OSC_{aug}$ used the $\mathbf{X}_{aug}$ and $\mathbf{y}_{aug}$ calibration base for calculating the orthogonal correction. The NA-NAP strategies performed NAP on the augmented calibration base $\mathbf{X}_{aug}$ and $\mathbf{y}_{aug}$ this way calculating a projection factor to transform the spectra in **X** to **X*** prior to PLS or CLS regression. The optimal number of factors $g_a$ to be removed was determined via the minimal RMSECV. The test set spectra underwent the same transformation as the calibration spectra.

*3.4.4. Strategies using ACLS*

The SRACLS procedure described in [14] was used to define an augmented pure component matrix **K** consisting of the pure spectra from the API (**A**) and the excipients (**B**), and the *p* first loading vectors from the residual matrix $\mathbf{E}_A$. In the PACLS procedure the augmented pure component matrix of the SRACLS procedure was further augmented with the *k* first loading vectors from **D**. The optimal *p* and *k* were determined through cross-validation (minimal RMSECV).

**3.5. Model update for strong water variations**

To assess whether the strategies using prior spectral information can account for strong water variations, 2 different model update strategies were proposed. In *model update 1*, the spectra of such processed tablets were added (with corresponding reference values) to the calibration set 1 to obtain an updated calibration set containing 197 samples (120 initial + 77 processed samples). The prior information contained, just as in the previous section, no spectra of processed tablets. In *model update 2*, the initial training set 1 (120 not-processed tablets) was used for calibration. The prior information matrix containing intra- and inter-batch variability was augmented with spectra from processed tablets to account also for the strong water variations.

## 4. Results and discussion

### *4.1. Models using prior information containing intra- and inter-batch variation*

In this section, the above discussed strategies were applied to evaluate their capability for filtering the intra- and inter-batch effects hampering the model performance. Test set 1 was used to evaluate the model's ability to withstand unknown variability (external validation). With PLS models covering different amounts of variation in their calibration set as a benchmark, the performance of the investigated modelling strategies was investigated in terms of number of latent variables (LV's), root mean square error of prediction (RMSEP) and its statistical significance (Table 1). The number of factors used for the correction method was also reported. All models were trained on 4 different sample sets, holding different amounts of variation. Set 1 (120 tablets, 12 batches) contained the most intra- and inter-batch variation, while set 4 (18 tablets, 6 batches) contained the least. Sets 2 (60 tablets, 6 batches) and 3 (30 tablets, 6 batches) contained more intra-batch variation compared to set 4. The prior information matrix **C** was considered to be representative for intra- and inter-batch variability in the spectra.

The PLS, NAP-CLS, NAP-PLS and OSC-PLS models did not use any prior information and were added for comparison purposes. They were marked in Table 1 with an asterisk. For all calibration sets the OSC-PLS, NAP-PLS and NAP-CLS models were generally more parsimonious (less LV's) compared to PLS, but did not show improved predictive power. Many papers present similar conclusions on this behavior of OSC and NAP [22-24]. PLS established higher prediction errors in models calibrated on sample sets containing less variability.

The prediction performances of the ACLS models (SRALCS and PACLS) were much worse than the PLS-based models, making the latter models more suitable for the present case study. It is hypothesized that ACLS models, based on a linear-additive model may have trouble capturing the physical differences between the NIR spectra of powders (pure components) and tablets. They might be more suitable for modelling pure mixtures of different components [14]. Augmenting the calibration set prior to PLS, i.e. by noise addition (NA)-PLS, yielded models with more LV's compared to the original PLS models. This can be attributed to more spectral variation being included in the calibration set. Noise augmentation before PLS regression (NA-PLS), making the calibration matrix very large (10213 objects) and noisy, could not significantly enhance the prediction performance either, nor could OSC when applied on this augmented calibration matrix (NA-OSC-PLS). NA-NAP-PLS showed a statistically better prediction performance than the PLS model ($p < 0.05$), with a 34-40% lower RMSEP, while needing less latent variables. A similar behavior of this approach was also seen on all other calibration sets with a progressively smaller size (Table 1).

Compared to NA-PLS, adding the NAP pre-processing seems to be advantageous for making the model more parsimonious and for improving the prediction performance. Nadler and Coifman [27] concluded that PLS may behave differently in two idealized settings: for a noise-free training set the regression vector computed by PLS is, up to normalization, the net analyte signal, while for a noisy infinite training set the regression vector is not purely proportional to the NAS vector, but is optimal under a mean squared error of prediction criterion. Considering the obtained results, it may be speculated that NAP may have a different relationship towards PLS in the augmented calibration set. It would be an interesting topic of further study to find out why the NAP algorithm performs so much better on the large and noisy augmented calibration sets, and actually allows here an improvement in prediction performance.

Compared to NAP-PLS, the NA procedure may add more variability into the calibration data base, which is missing when using non-exhaustive calibration. In the NAP-CLS and NAP-PLS models only 7 to 8 factors were removed by the NAP procedure, whereas the NA-NAP operation removed between 10 and 12 factors, and approximately 1-3 additional factors were needed for subsequent PLS regression. This may approximate the number of perturbation factors to be expected in future tablets. Hence the augmentation step is also necessary to make the model more robust and the completeness of the clutter is an important factor for the performance and robustness of the models.

Except for the model calibrated on set 4 the NA-NAP method performed clearly better when combined with PLS instead of CLS. In these cases, PLS was able to further model the important information. In case of calibration set 4, the NA-NAP operation returned the original calibration spectra **X** into **X*** containing almost exclusively information on the component to be quantified (Fig.8), with PLS and CLS giving a comparable result.

Compared to PLS, EPO-PLS based calibrations showed lower RMSEP's and needed a lower number of LV's, and especially for models calibrated on sample sets carrying fewer variability there was a statistical difference between both approaches. Considering only the chemical variations in the correction, i.e. $EPO_{chem}$, allowed improving the model performance to some extent. This can be attributed to the correction for the changing tablet compositions over the API target concentration levels, as well as for excipient inhomogenities among individual tablets.

*4.2. Deliberate water variation (model update)*

When the tablets are processed, i.e. stored at different extreme temperature and humidity conditions, the water content in the tablets can change considerably [6]. Compared to the tablets that were stored under normal conditions, those stored long in high humidity conditions (i.e. 75% RH) showed an interesting effect in their NIR spectra, i.e. shifts in the API absorption signals were

visible (Fig.2). These frequency shifts may suggest a different hydrogen bonding state of the API [28]. As this variation affects the PLS model predictions (Table 2), we investigated whether this information might be considered in the calibration base or in the clutter for a model update.

In case the model was updated by adding deliberate water variation to the calibration data base, the NA-NAP-PLS strategy again outperformed the others, and was the only model being statistically significantly better than PLS (Table 2, update 1). With a model requiring only 4 LV's an RMSEP of 0,380 for test set 1 and one of 0,385 for test set 2 were obtained. Hence the model was able to predict perturbed samples reasonably well (test set 2), while also not substantially reducing the prediction performance for normal or unperturbed samples (test set 1). Compared to NA-NAP-CLS, NA-NAP-PLS allowed a lower RMSEP for both test sets. The NA-NAP-PLS model now required 1 LV more than before model update (Table 1, set 1), which it probably needs to model the shifting of the API absorption signals.

The second model updating strategy included the spectra of the perturbed samples in the prior spectral information matrix. As expected, models using no prior information from processed tablets, i.e. PLS, NAP-CLS, NAP-PLS, OSC-PLS, SRACLS and $EPO_{chem}$-PLS, displayed poor prediction performance for test set 2 (Table 2, update 2). This could be attributed to the fact that they did not take into account the deliberate water variation. Although some of the modelling strategies that used prior information from processed tablets were able to get better prediction errors for test set 2, the overall model performance was low and none of the methods performed significantly better than PLS. Compared to the models updated with strategy 1, none of the methods in model update 2 was able to obtain good prediction results for test set 2. The prediction performance for test set 1 remained reasonably well in all models (except for PACLS), whereas predictions for test set 2 were always poor. The reason for this was found to be that the spectral perturbations due to strong water content variations in C (also causing shifting of the API absorption signals) are not completely independent of **y**, and therefore could not be considered in the model update. The proposed method to define the clutter (step 1 of the strategy, see section 4.1.1) uses unsupervised orthogonal projections with the pure component spectra. These may remove that part of **C**, which is co-linear to **y**, making it not possible to fully consider this kind of spectral variation in the clutter. These results demonstrate that the proposed strategies only allow making the model more robust for spectral variation that is not co-linear with the one from the property of interest. Hence, this variability should still be considered in the calibration set.

## 5. Conclusions

When developing a NIR model for routine quality control that needs to operate over a longer time course, it may not be evident to cover all possible perturbation factors in the calibration set.

Because it is difficult to select samples describing intra-and inter-batch variability and to decide when the calibration set is representative, this study explored whether the judicious use of such prior spectral information during model development can improve the performance of NIR models for predicting the API content in tablets. The proposed strategies did neither require exact knowledge of the perturbation levels (only controlled variability of the considered perturbations was necessary), nor needed extra reference analyses.

Rather than exhaustive calibration, a more cost-effective model development approach was aspired. A massive amount of prior spectral information on intra- and inter-batch variation was obtained to allow a more representative view of the possible disturbing (i.e. systematic) effects to be encountered in the NIR spectra of tablets from a routine production. The best approach (NA-NAP-PLS) consists of four essential steps. First, the disturbances to be removed are identified by means of using an orthogonal projection of pure component spectra (API and excipients) to those of the tablets containing representative intra- and inter-batch variability. This way, the known chemical variation is removed from the spectra leaving only the other variations. These mean-centered variations are in a second step added to *n* repetitions of the spectra of **X** for calibration (noise augmentation) to increase the variability in the existing calibration database. In a third step, net analyte pre-processing is used to remove those variations in $\mathbf{X}_{aug}$ that are orthogonal to $\mathbf{y}_{aug}$, in order to obtain a better estimate of the NAS. Finally, in a fourth step the orthogonally corrected **X** spectra are regressed to **y** by PLS regression.

Compared to PLS, a statistically significant improvement in prediction performance and a 34-40% reduction in RMSEP was obtained for predicting new tablets with unknown intra- and inter-batch variability. The model needed a minimal amount of calibration samples and latent variables. It would be interesting to further study the behavior of NAP-PLS in very large and noisy calibration matrices.

Where PLS models need the calibration set to be as complete as possible, models using prior spectral information require completeness of their clutter. This was also demonstrated by updating the model for strong water content variations in the tablets. The latter caused frequency shifts of the API absorption signals that could not be included in the clutter, resulting in reduced performance of the proposed strategies for this kind of variation. It also revealed a limitation of the method: the spectral variance that is co-linear with the wanted variation could not be considered in the clutter and should be considered in the calibration set.

**Acknowledgements**

Sigrid Pieters and Wouter Saeys are respectively funded as aspirant and postdoctoral fellow of the Research Foundation – Flanders (FWO).

**References**

[1] J. Luypaert, D.L. Massart, Y. Vander Heyden, Talanta 72 (2007) 865-883.

[2] T. De Beer, A. Burggraeve, M. Fonteyne, L. Saerens, J.P. Remon, C. Vervaet, Int. J. Pharm. 147 (2011) 32-47.

[3] D. Cozzolino, W. Cynkar, N. Shah, P. Smith, Anal. Bioanal. Chem. 401 (2011) 1475-1484.

[4] W.F. McClure, Anal. Chem. 66 (1994), 43-53.

[5] M. Blanco, A. Peguero, J. Pharm. Biom. Anal. 52 (2010) 59-65.

[6] T. Van den Kerkhof, R. De Maesschalck, K. Vanhoutte, M.C. Coene, J. Pharm. Biomed. Anal. 42 (2006) 517-522.

[7] A. Peirs, J. Tirry, B. Verlinden, P. Darius, B.M. Nicolaï, Postharv. Biol. Technol. 28 (2003) 269-280.

[8] EMEA/CHMP/CVMP/QWP/17760/2009 Rev 2 (2012), accessed on 28/03/2012

[9] D.L. Massart, B.M.G. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of chemometrics and qualimetrics: part A, Elsevier, Amsterdam, 1997.

[10] M.J. Saiz-Abajo, B.H. Mevik, V.H. Segtnan, T. Naes, Anal. Chim. Acta 533 (2005) 147-159.

[11] B.H. Mevik, V.H. Segtnan, T. Naes, J. Chemom. 18 (2004) 498-507.

[12] J.A. Fernandez Pierna, F. Chauchard, S. Preys, J.M. Roger, O. Galtier, V. Baeten, P. Dardenne, Chemom. Intell. Lab. Syst. 106 (2011) 152-159.

[13] V.H. Segtnan, B.H. Mevik, T. Isaksson, T. Naes, Appl. Spectr. 59 (2005) 816-825.

[14] W. Saeys, K. Beullens, J. Lammertyn, H. Ramon, T. Naes, Anal. Chem. 80 (2008) 4951-4959.

[15] D. M. Haaland, D.K. Melgaard, Appl. Spectr. 54 (2000) 1303-1312.

[16] D.M. Haaland, D.K. Melgaard, Vib.Spec. 29 (2002) 171-175.

[17] J.C. Boulet, J.M. Roger, Chemom. Intell. Lab. Syst., in press

[18] J.C. Boulet, T. Doco, J.M. Roger, Chemom. Intell. Lab. Syst. 87 (2007) 295-302.

[19] J.M. Roger, F. Chauchard, V. Bellon-Maurel, Chemom. Intell. Lab. Syst. 66 (2003) 191-204.

[20] S. Preys, J.M. Roger, J.C. Boulet, Chemom. Intell. Lab. Syst. 91 (2008) 28-33.

[21] S. Wold, H. Antti, F. Lindgren, J. Ohman, Chemom. Intell. Lab. Syst. 44 (1998) 175-185.

[22] H. C. Goicoechea, A. C. Olivieri, Chemom. Intell. Lab. Syst. 56 (2001) 73-81.

[23] W. Ni, S.D. Brown, R. Man, Chemom. Intell. Lab. Syst. 98 (2009) 97-107.

[24] O. Svensson, T. Kourti, J.F. MacGregor, J. Chemom. 16 (2002) 176-188.

[25] H.R. Cederkvist, A.H. Aastveit, T. Naes, J. Chemom. 19 (2005) 500-509.

[26] http://wiki.eigenvector.com/index.php?title=Declutter Settings Window, accessed on 11/04/2012.

[27] B. Nadler, R.R. Coifman, J. Chemom. 19 (2005) 45-54.

[28] S. Pieters, T. De Beer, J.C. Kasper, D. Boulpaep, O. Waszkiewicz, M. Goodarzi, C. Tistaert, W. Friess, J.P. Remon, C. Vervaet, Y. Vander Heyden, Anal. Chem. 84 (2012) 947-955.

**Figure captions**

**Fig. 1.** NIR spectra of the pure tablet components (SNV pre-processed) in the range 9000-7600 $cm^{-1}$.

**Fig. 2.** NIR spectra of processed and unprocessed tablets (SNV pre-processed) in the range 9000-7600 $cm^{-1}$. Spectra from unprocessed tablets are in grey, from tablets stored for 30 min at 50°C in green, from tablets stored for 3h and 16h at 75% RH in purple and red, respectively. The black ellipses indicate visible shifting of the API absorption signals.

**Fig. 3.** PCA score plot (PC1 versus PC2) from **C** covering the 10213 spectra of tablets from 27 different batches (in grey). The spectra used in the calibration set 1 are represented by differently coloured markers according to their batch.

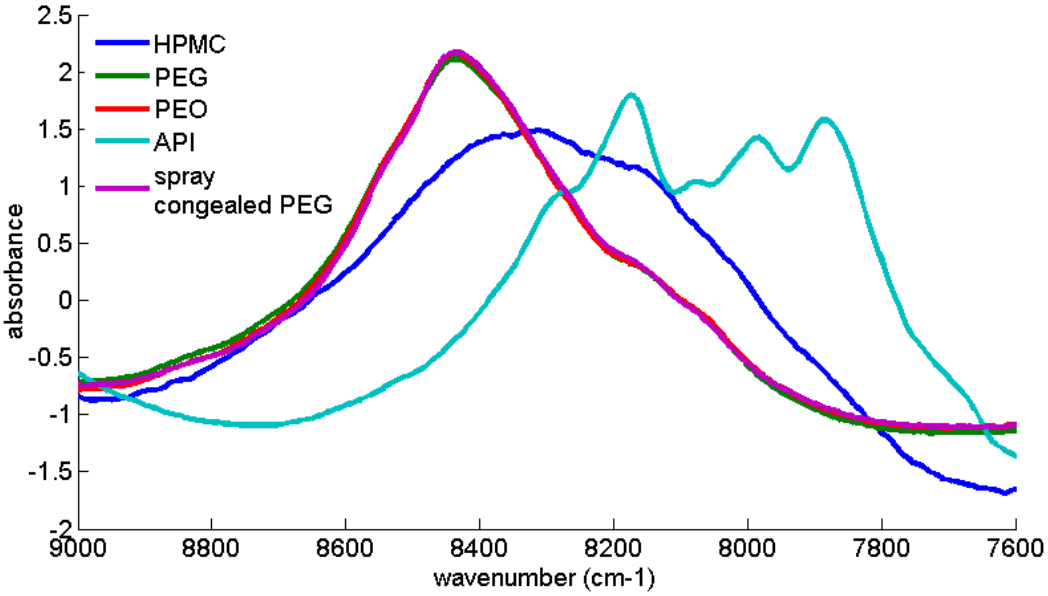
**Fig. 4.** PCA score plot (PC1-PC2-PC3) of NIR spectra (SNV pre-processed) from tablets with 25 mg target API concentration manufactured in a production line. The different batches are indicated with differently colored markers.

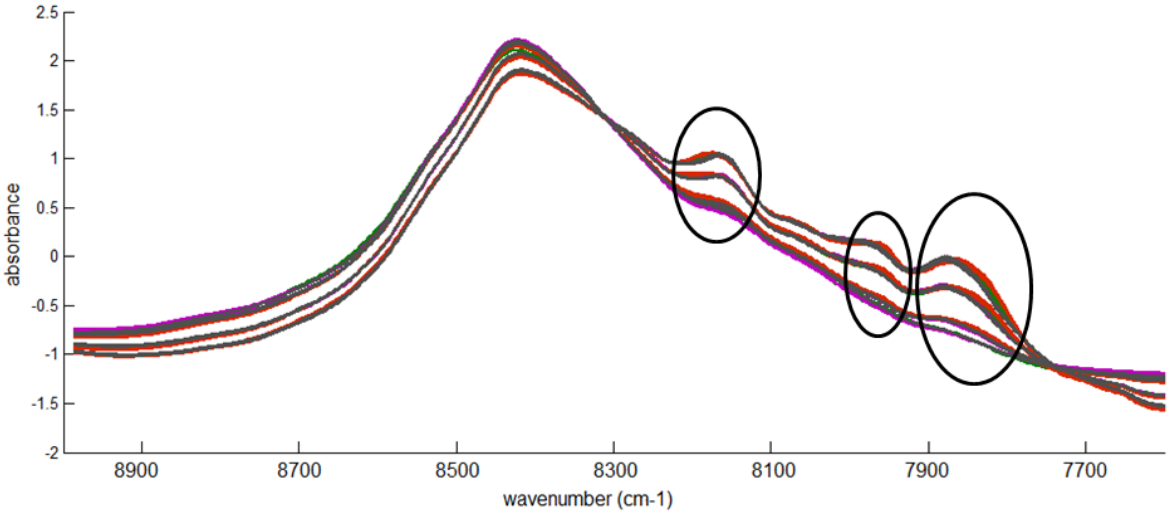
**Fig. 5.** PCA score plot (PC1-PC2-PC3) of NIR spectra (SNV pre-processed) from tablets with 100 mg target API concentration manufactured in a production line. The different batches are indicated with differently colored markers.

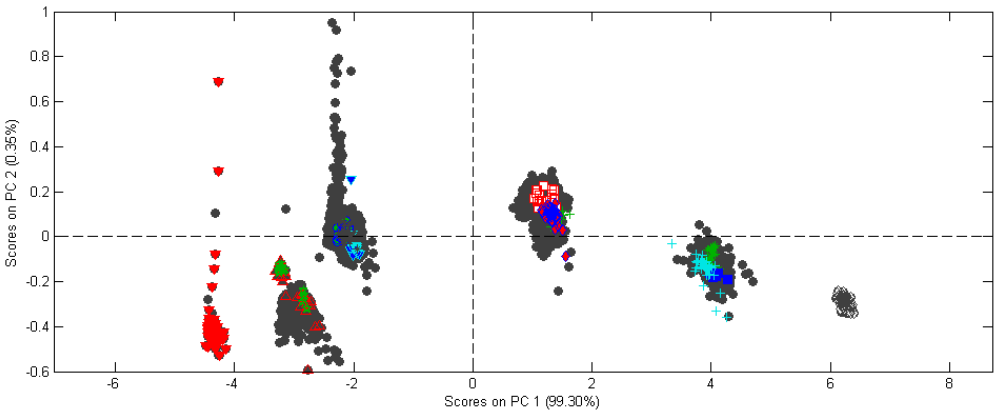
**Fig.6.** Spectra before (A) and after (B) orthogonal correction using the pure component spectra of the API and excipients to remove the known chemical variations.

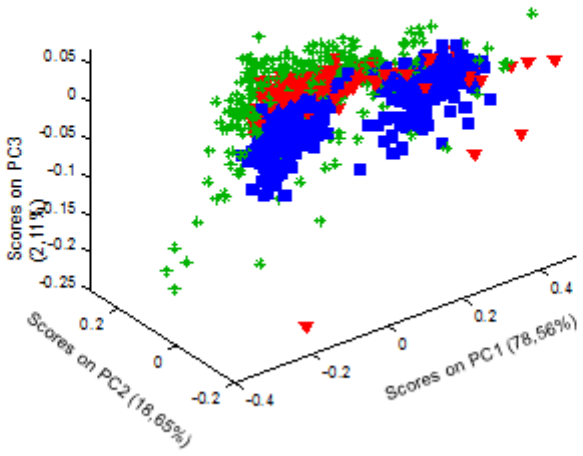
**Fig.7.** Flow chart depicting the applied methodologies. **A, B** and **C** are matrices containing prior information; the subscript *c* represents mean-centered matrices. **X** (measured spectral absorbances **J**) and **y** (concentrations to be predicted **P**) represent the original calibration set. $\mathbf{X}_{aug}$ and $\mathbf{y}_{aug}$ are the noise augmented calibration set. $P^{\perp}$ represents orthogonal projection and **X*** is the calibration matrix returned after correction. **K** is the matrix of pure component signals at unit concentration.

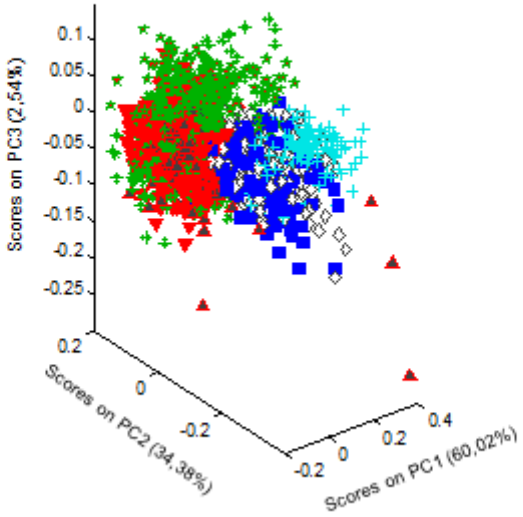
**Fig. 8.** Calibration spectra $\mathbf{X}_{nap}$***** (18 x 364) of set 4 after NA-NAP transformation. These spectra were subsequently calibrated on **y** with PLS. The different colors represent the different API concentration levels.

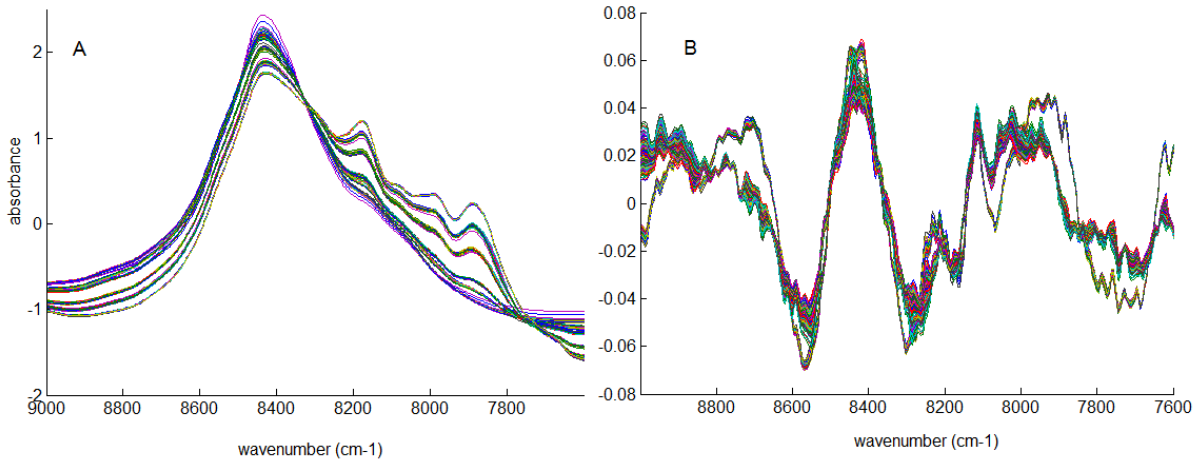

HPMC
PEG
PEO
API
spray
congealed PEG
absorbance
wavenumber (cm-1)
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
9000
8800
8600
8400
8200
8000
7800
7600

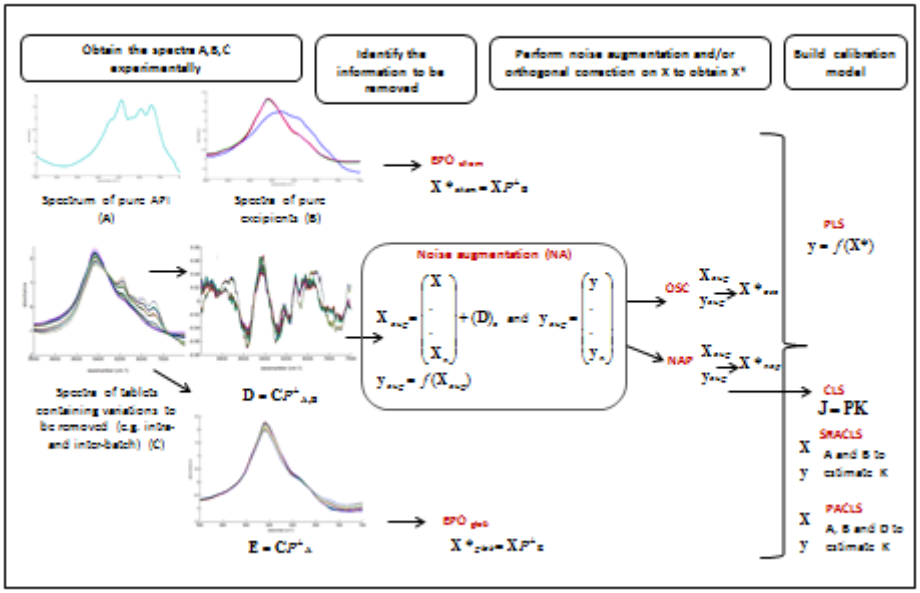

absorbance
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
8900
8700
8500
8300
8100
7900
7700
wavenumber (cm-1)


Scores on PC 2 (0.35%)
Scores on PC 1 (99.30%)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-6
-4
-2
0
2
4
6
8


Scores on PC3 (2,11%)
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.25
Scores on PC2 (18,65%)
0.2
0
-0.2
Scores on PC1 (78,56%)
-0.4
-0.2
0
0.2
0.4


Scores on PC3 (2,54%)
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.25
0.2
0
-0.2
Scores on PC2 (34,38%)
-0.2
0
0.2
0.4
Scores on PC1 (60,02%)


A
absorbance
2
1
0
-1
-2
9000
8800
8600
8400
8200
8000
7800
wavenumber (cm-1)
B
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
8800
8600
8400
8200
8000
7800
7600
wavenumber (cm-1)


Obtain the spectra A,B,C experimentally
Identify the information to be removed
Perform noise augmentation and/or orthogonal correction on X to obtain X*
Build calibration model
Spectrum of pure API (A)
Spectra of pure excipients (B)
EPO alum
X*alum = XP+B
Spectra of tablets containing variations to be removed (e.g. intra- and inter-batch) (C)
D = CP+A,B
Noise augmentation (NA)
Xaug = (X ... Xn) + (D)n and yaug = (y ... yn)
yaug = f(Xaug)
OSC
NAP
E = CP+A
EPO glob
X*glob = XP+E
PLS
y = f(X*)
CLS
J = PK
SRACLS
A and B to estimate K
PACLS
A, B and D to estimate K

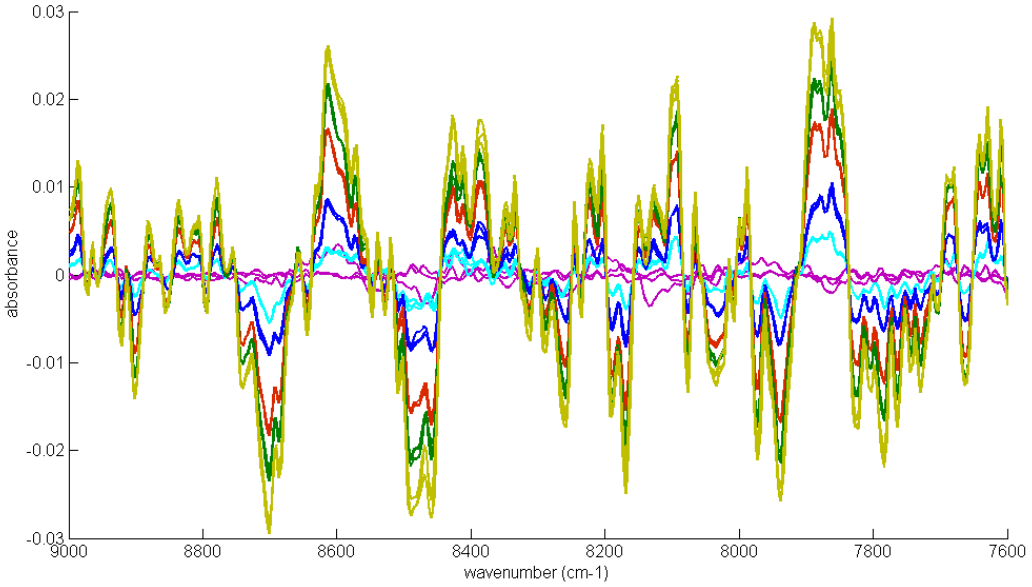

0.03
0.02
0.01
0
-0.01
-0.02
-0.03
absorbance
9000
8800
8600
8400
8200
8000
7800
7600
wavenumber (cm-1)

**Table 1.** Prediction performance (for test set 1) of the different models calibrated on sample sets with different variation. The matrix with prior intra- and inter-batch variation was **C** (10213 x 364). The best prediction results are in bold.

| Calibration set | Set 1 (120 tablets, 12 batches) | | | | Set 2 (60 tablets, 6 batches) | | | | Set 3 (30 tablets, 6 batches) | | | | Set 4 (18 tablets, 6 batches) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **# factors** | **# LV** | **RMSEP** | ***p* value[b]** | **# factors** | **# LV** | **RMSEP** | ***p* value[b]** | **# factors** | **# LV** | **RMSEP** | ***p* value[b]** | **# factors** | **# LV** | **RMSEP** | ***p* value[b]** |
| PLS[a] | - | 7 | 0.586 | - | - | 4 | 0.630 | - | - | 4 | 0.636 | - | - | 4 | 0.663 | - |
| NAP-CLS[a] | 7 | - | 0.547 | 0.849 | 8 | - | 0.739 | 0.543 | 7 | - | 0.755 | 0.512 | 7 | - | 1.126 | <0.001[c] |
| NAP-PLS[a] | 7 | 4 | 0.601 | 1.000 | 8 | 5 | 0.608 | 1.000 | 7 | 4 | 0.604 | 0.807 | 7 | 4 | 1.123 | <0.001[c] |
| OSC-PLS[a] | 3 | 4 | 0.560 | 1.000 | 3 | 5 | 0.646 | 0.986 | 1 | 3 | 0.636 | 1.000 | 1 | 3 | 0.663 | 0.999 |
| NA-PLS | - | 8 | 0.496 | 0.928 | - | 5 | 0.588 | 1.000 | - | 6 | 0.566 | 0.963 | - | 6 | 0.585 | 0.985 |
| SRACLS | 5 | - | 0.659 | 0.994 | 10 | - | 1.254 | <0.001[c] | 6 | - | 0.975 | <0.001[c] | 6 | - | 0.988 | <0.001[c] |
| PACLS | 5-1 | - | 0.908 | <0.001[c] | 8-1 | - | 1.099 | <0.001[c] | 6-1 | | 0.945 | <0.001[c] | 6-1 | - | 1.139 | <0.001[c] |
| $EPO_{chem}$-PLS | 4 | 5 | 0.511 | 0.716 | 4 | 4 | 0.468 | 0.013[c] | 4 | 5 | 0.487 | 0.107 | 4 | 5 | 0.519 | 0.281 |
| $EPO_{gl}$-PLS | 1 | 6 | 0.461 | 0.449 | 1 | 7 | 0.581 | 1.000 | 1 | 6 | 0.478 | 0.004[c] | 1 | 7 | 0.516 | 0.009[c] |
| NA-NAP-CLS | 10 | - | 0.572 | 1.000 | 12 | - | 0.527 | 0.588 | 11 | - | 0.546 | 0.370 | 11 | - | 0.434 | 0.026[c] |
| **NA-NAP-PLS** | **10** | **3** | **0.388** | **0.027[c]** | **12** | **2** | **0.407** | **0.007[c]** | **11** | **2** | **0.382** | <0.001[c] | **11** | **1** | **0.431** | **0.015[c]** |
| NA-OSC-PLS | 2 | 3 | 0.488 | 0.697 | 1 | 2 | 0.546 | 0.586 | 1 | 2 | 0.533 | 0.424 | 1 | 2 | 0.564 | 0.727 |

[a] These models did not use prior information; they used only information from the calibration set

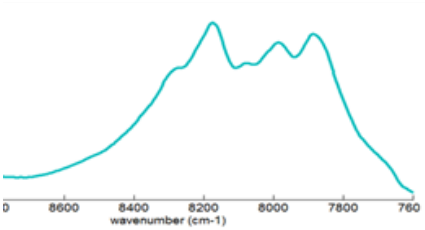
[b] *p* values (Dunnett's test) for the significance testing by two-way ANOVA of the predictive ability compared to PLS

[c] Indicates significantly lower absolute prediction errors on significance level $\alpha$=0.05

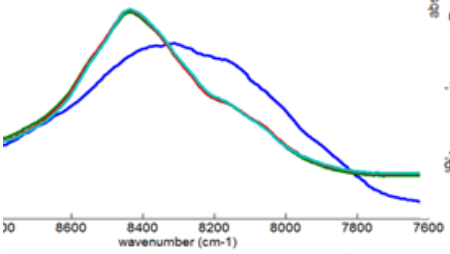
**Table 2.** Prediction performance (for test sets 1 and 2) after model update for deliberate water variations. The best prediction results are in bold.

| Model | Update 1 | | | | | Update 2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **# factors** | **# LV** | **RMSEP test set 1** | **RMSEP test set 2** | ***p* value[b]** | **# factors** | **# LV** | **RMSEP test set 1** | **RMSEP test set 2** | ***p* value[b]** |
| PLS[a] | - | 8 | 0.632 | 0.407 | - | - | 7 | **0.586** | 1.274 | - |
| NAP-CLS[a] | 8 | - | 0.534 | 0.461 | 0.883 | 7 | - | 0.547 | 1.488 | 1.000 |
| NAP-PLS[a] | 8 | 6 | 0.644 | 0.302 | 0.645 | 7 | 4 | 0.601 | 1.358 | 1.000 |
| OSC-PLS[a] | 5 | 6 | 0.632 | 0.346 | 0.989 | 3 | 4 | 0.560 | 1.506 | 1.000 |
| NA-PLS | - | 8 | 0.548 | 0.433 | 0.549 | - | 5 | 0.466 | 0.854 | 0.844 |
| SRACLS[a] | 9 | - | 0.780 | 0.466 | 0.124 | 5 | - | 0.659 | 1.557 | 1.000 |
| PACLS | 9-1 | - | 0.756 | 0.417 | 0.067 | 5-12 | - | 5.490 | 8.323 | <0.001[c] |
| $EPO_{chem}$-PLS[a] | 4 | 5 | 0.575 | 0.376 | 0.792 | 4 | 5 | 0.511 | 0.913 | 0.948 |
| $EPO_{gl}$-PLS | 1 | 7 | 0.516 | 0.397 | 0.284 | 1 | 7 | 0.549 | 0.977 | 0.989 |
| NA-NAP-CLS | 11 | - | 0.512 | 0.517 | 0.964 | 13 | - | 0.567 | 0.728 | 0.923 |
| NA-NAP-PLS | **11** | **4** | **0.380** | **0.385** | **0.003[c]** | 13 | 3 | 0.488 | 0.761 | 0.595 |
| NA-OSC-PLS | 4 | 2 | 0.498 | 0.409 | 0.265 | 1 | 2 | 0.510 | 0.976 | 0.966 |

[a]These models did not use prior information of processed tablets containing deliberate water variations. [b]*p* values (Dunnett's test) for the significance testing by two-way ANOVA of the predictive ability compared to PLS. [c]Indicates significantly lower absolute prediction errors on significance level α=0.05.

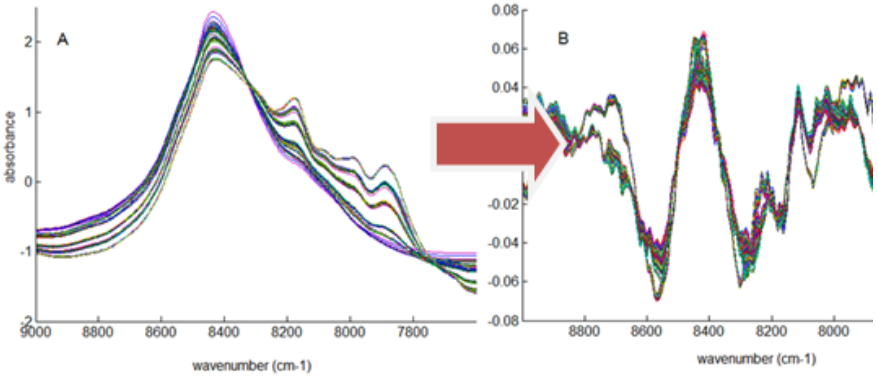



(A) pure API




(B) pure excipients

$$\mathbf{C}P^{\perp}_{\mathbf{A,\,B}}$$




(C) spectra from tablets with variations to be removed

(D) clutter

**Highlights:**

- a cost-effective NIR model development strategy is proposed for API content prediction in tablets

- judicious use of prior spectral information improves PLS model performance

- the clutter captures representative intra- and inter-batch spectral variability to be removed

- model requires completeness of the clutter rather than comprehensive calibration sets